Entity Popularity on the Web: Correlating ANSA News and AOL Search

Angela Fogarolli, George Giannakopoulos, and Heiko Stoermer

University of Trento,
Dept. of Information Science and Engineering,
Via Sommarive 14, 38123 Trento, Italy
{afogarol,ggianna,stoermer}@disi.unitn.it

Abstract. Annotated Web content, digital libraries, news and media portals, e-commerce web sites, online catalogs, RDF/OWL knowledge bases and online encyclopedias can be considered containers of *named entities* such as organizations, persons, locations. Entities are mostly implicitly mentioned in texts or multi-media content, but increasingly explicit in structured annotations such as the ones provided by the Semantic Web. Today, as a result of different research projects and commercial initiatives, systems deal with massive amounts of data that are either explicitly or implicitly related to entities, which have to managed in an efficient way. This paper contributes to Web Science by attempting to measure and interpret trends of entity popularity on the WWW, taking into consideration the occurrence of named entities in a large news corpus, and correlating these findings with analysis results on how entities are searched for, based on a large search engine query log. The study shows that entity popularity follows well-known trends, which can be of interest for several aspects in the development of services and applications on the WWW that deal with larger amounts of data about (named) entities.

1 Introduction

Entities – as opposed to documents – are gaining importance in the information space of the Web, due to the many activities that attempt to elevate the Web from a document-centric information space that is very hard to process for machine algorithms, towards something more fact-centric, powered by machine readable annotations. At the core lie (real-world) entities such as persons, locations, organizations, which often represent the first-class objects that facts are *about*.

It is entities mentioned on the web that this paper is concentrating on. Information systems that do not operate on closed, controlled environments, but which deal with information about entities on the web, often need to implement methods that rely on a measure of popularity to take operational decisions. These can relate to cache management (data about which entity do I maintain in the cache), data lifecycle management (are my data still up-to-date), improvement

D. Dicheva and D. Dochev (Eds.): AIMSA 2010, LNAI 6304, pp. 141–150, 2010.

of search results (in doubt, return the more popular entity as a higher-ranked result), and other use cases that we cannot yet foresee.

Measuring popularity of entities mentioned on the web however is a challenge. For an enormous information space like the Web, standard approaches like simple statistical measures seem hardly practical. Running the required Named Entity Extraction (NER) and entity consolidation methods on the web itself can – if at all – only be performed by a chosen few, which leads to the situation that to the best of our knowledge, no study about the behavior of entity popularity on the Web is available today.

To tackle the challenges posed by the sheer size of the Web, this paper attempts to derive facts from the analysis of a large, yet manageably sized corpus of news items, develop a hypothesis, and confirm this hypothesis by correlating it to the analysis of a large search engine query log. Following this approach, we hope to provide results that allow for a certain level of generalization, and thus create insights about the behavior of entity popularity on the Web itself.

The rest of this article is arranged as follows. First, in Sect. 2 we provide a working definition of what we consider to be an "entity". Then we give a short overview over related work in section 3. Section 4 presents the analysis performed using the Italian news agency (ANSA) Annotated Corpus, and develops and confirms a hypothesis with the help of the AOL search engine query log [16]. Sect. 5 discusses the work performed, and concludes the article, offering use-cases in which the results presented can create a benefit, and illustrating future work.

2 Definition of Entity

While the Semantic Web works with the notion of a "resource", which is basically anything that is identified by a URI, concepts, objects and properties alike, our interest is more similar to what traditional Description Logics knowledge bases call "individuals". A straight-forward and trivial definition of the notion of "entity" in our sense might be the idea of a "real-world object", but already the example of a geographical location such as a city leads to first complications (is a city an object?). And indeed, a definition that describes our intent and at the same time stands up acceptably well to philosophical requirements is surprisingly difficult to find. Several important aspects have to be considered, such as physical vs. abstract, spatial vs. temporal, individual vs. collection, behavior in time, agentivity, and more. Bazzanella has come up with working solution that covers our area of interest, being the *set of all particular and concrete entities (events included)* [1]. This means that classes, properties and abstract concepts do not count as entity in our sense, but people, organizations, and geographical locations do.

3 Related Work

Extracting entities from the news is a current trend as stated by the the International Press Telecommunication Council(IPTC) [8]:

> Increasingly, news organizations are using entity extraction engines to find "things" mentioned in news objects. The results of these automated processes may be checked and refined by journalists. The goal is to classify news as richly as possible and to identify people, organizations, places and other entities before sending it to customers, in order to increase its value and usefulness.

The question which *trends* entity popularity follows – especially on the Web – has not been widely analyzed so far. There is related work from the areas of focused retrieval [5,14,15,9] or question answering [4,17], which mainly describes techniques to measure entity popularity on a given corpus to achieve certain goals. First studies on social networks on the web [12,11] deal with inferring a ranking of entities based on their status in social networks. None of these works however intend to generalize and create insights about the Web itself.

4 A Study on Entity Popularity

4.1 Entities in ANSA News: The Italian News Agency Corpus

When selecting a set of documents that would allow to make observations suitable for a generalization to the Web as a whole, several aspects need to be covered. First, the dynamic character of the Web should be reflected, meaning that for example a collection of literature texts would not be an optimal choice. Secondly, the corpus should cover a rather broad area of domains and not focus on a single topic. Third, it should reflect the fact that the Web often covers many topics of broad interest, especially related to *entities* according to our definition.

The ANSA[1] Annotated Corpus is very useful in this respect, as it fulfills these criteria: it contains almost 1.2 Million news articles from a period of 3 years, about everything that is new and noteworthy, including persons, organizations, and locations.

The annotation of the ANSA corpus is made by analyzing the news items using Name Entity Recognition (NER). NER is composed by an Entity Detection phase which parses the text and determine the entities, and an Entity Normalization phase which associates a detected entity with its identifier. The NER aims at the detection of people, organizations, locations and events inside the news articles. While Entity Detection is a stand-alone functionality completely developed by Expert System[2], for Entity Normalization (to provide entities with their ids) the Entity Name System (ENS) [3] is queried to retrieve unique identifiers for each detected entity.

[1] "ANSA (Agenzia Nazionale Stampa Associata), is the leading wire service in Italy, and one of the leaders among world news agencies. ANSA is a not-for-profit cooperative, whose mission is the distribution of fair and objective news reporting. ANSA covers national and international events through its 22 offices in Italy, and its presence in more than 80 cities in 74 countries in the world." (cf. `http://en.wikipedia.org/wiki/Agenzia_Nazionale_Stampa_Associata`)

[2] `http://www.expertsystem.net/`

Table 1. Number of distinct and total entities in the ANSA corpus, by entity type

Entity type	Abbr.	Distinct E.	Total E.
Location	LOC	33409	12474210
Organization	ORG	77888	847889
Person	PER	138143	4616958
Total		249440	17939057

The size of the corpus and metadata ensure that the trends and numbers we describe are statistically reliable. Table 1 gives more details about the number of entities contained. The entities in the corpus are classified into three categories: person(PER), organization(ORG) and location(LOC).

Formally, for our later analysis, we model an entity e as follows: $e = \langle l, t \rangle$ with string label l and type t, where $t = \{LOC, PER, ORG, AUT\}$

An article A is modeled as: $A = \langle E, d, id, C \rangle$ with $E = \{e\}$, the set of entities mentioned in A, the publication date d, the unique article ID (URL) id, and the set of classifiers C from several formal taxonomies.

Entity Lifespan. An important aspect we are interested in to gain a better understanding of entity behavior is the "lifespan". We define lifespan as the number of months between the first and last occurrence of an entity. Table 2 shows the lifespan quantiles for each entity type.

Table 2. Quantiles of entity lifespan in the ANSA corpus in months

Entity type	Min	1st Q.	Median	Mean	3rd Q.	Max
LOC	2.00	17.00	29.00	25.12	35.00	36.00
ORG	2.00	6.00	14.00	16.06	25.00	36.00
PER	2.00	8.00	18.00	18.17	28.00	36.00
Overall	2.00	9.00	20.00	19.46	30.00	36.00

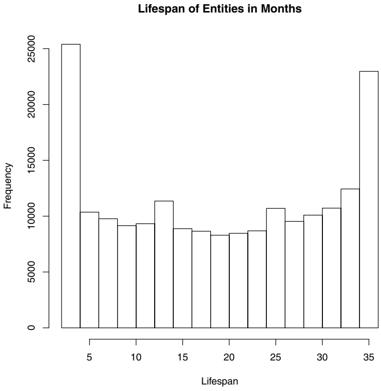

Fig. 1. The NE lifespan histogram

We call "NE lifespan" the average lifespan of the entities for the general named entity types: person, organization and location. The NE lifespan average for the entities in the corpus is 17.89 months. In Figure 1 we illustrate the histogram of the NE lifespan.

Entity Re-use. Looking closer at the data, it is possible to understand trends and behavioral patterns of entity popularity. This analysis is focused on learning how many times individual entities are mentioned in the corpus, i.e. how many occurrences inside the corpus exist for the same entity. From a different perspective, multiple occurrences of the same entity can be considered "re-use" of the entity. We analyze how many of times an entity is mentioned per month. In table 3 we summarize the average values of re-use for each entity type. The re-use is calculated dividing the total occurrences by the distinct ones for each entity type. The overall re-use is the average re-use rate on the three years of news covered by the corpus.

Table 3. Simple overall average re-use rate per entity type

Entity type	Re-use rate
LOC	373.38
ORG	53.66
PER	33.42

4.2 Hypothesis and Evaluation

In this section we evaluate our hypothesis about entity popularity and its trends. The section is divided in two parts, the first one correlates annotations about entities in the news corpus with searched entities in a search engine query log, while the second describes the analysis made to confirm that entity popularity follows a power-law distribution.

Entity Popularity. It can be argued that the popularity of an entity on the Web can not be measured by means of the number of times it is mentioned in a news article, but the popularity is more related to how many times Internet users search for that particular entity. In order to show that the entities mentioned in the news can however indeed give an idea of popularity of entities in the WWW, in the following experiment we correlate a news corpus with a search engine log, with the aim of confirming findings from the one with analysis of the other.

Due to the fact that search log files are not widely available, we restricted this evaluation to a the query log of the AOL (American On Line) search engine, which covers a period of three months. We compare the entities in the AOL logs with the ANSA entities mentioned in the same period. For the sake of the evaluation, to prove the generality of the "popularity" aspect, we contextualize this analysis on the entity type *person*.

Table 4. Correlation test results. Statistically significance indicated in **bold** letters

Test	Correlation Value	P-value
Pearson	**0.42**	0.0
Spearman	**0.53**	0.0
Kendall	**0.41**	0.0

The AOL logs cover the period from March 1st, 2006 to May 31st, 2006. In order to correlate the entities in the logs with the ones in the news article metadata, we processed the AOL log using a Lucene[3] index. Next, we used Lucene index functionalities such as specific filters[4] to count the occurrences of the ANSA entities in the search logs.

The hypothesis tested is that *the popularity of an entity in the news behaves similarly to the popularity of an entity in search queries.* We expect to find that people increase the number of searches for a particular entity when the entity appears more often in the news. The hypothesis can be written like this:

> The number of occurrences of an entity in the news is correlated to occurrences of an entity in search queries over the same period.

We ran three different correlation experiments on our datasets using: Pearson, Spearman and Kendall correlation. In all three correlation tests, the p-value is used to indicate the statistical significance of the observation. The lower the p-value, the more "significant" the result, in the sense of statistical significance. As in most statistical analyses, a p-value of 0.05 or less indicates high statistical significance. The person entities appearing in both the ANSA (news) and AOL (query) datasets are 2745.

The Pearson's correlation indicates whether there exists a linear correlation between two random variables and measures the strength of this correlation. A Pearson correlation value close to 0 indicates that there is not linear correlation between the random variables. A value near -1 or 1, indicates full negative or positive linear correlation, i.e., the value of one variable is a linear function of the other. In correlating the AOL dataset with the ANSA subset we learned that *there is a very powerful and statistically significant linear correlation between the variables.*

Spearman's and Kendall's rank correlation indicate whether two random variables tend to be modified in a correlated manner, without taking into account whether the correlation is linear. In other words, if the value of one random variable is increased when the other is increased, the rank correlations will show

[3] Lucene: http://lucene.apache.org/

[4] For entity identification we apply different Lucene techniques which include the use of a PhraseQuery for matching the query string in the logs containing a particular sequence of terms. PhraseQuery uses positional information of the term that is stored in an index. The distance between the terms that are considered to be matched is called slop. With used a slop set to three to ensure that all name and last name of the entity in the ANSA is present in the log.

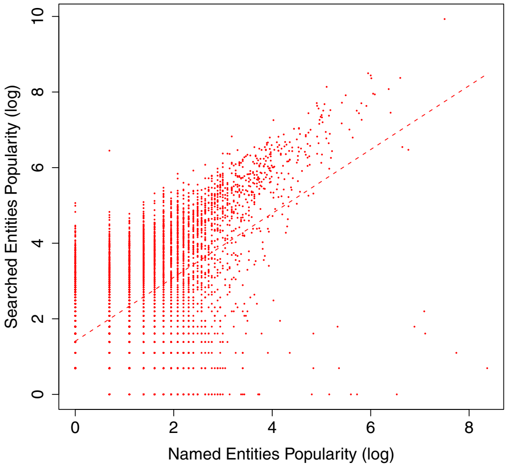

Fig. 2. Correlation between AOL searched entities and ANSA named entities (log-scale)

a value away from 0. The return value of the function describes the strength and direction (negative or positive) of the relationship between two variables. The results are between 1 and −1. The result of the rank coefficients indicate with a very high probability (since the p-value is almost 0.0) that there *is* a statistically significant correlation between entities in the AOL search logs and in the ANSA corpus. The intensity of the correlation is average.

The values of correlation mean that, if an entity is popular (i.e., appears often) in the news (ANSA corpus), it is expected to also be popular in the searches the users make and vice versa. In other words, each measure is a good predictor of the other over that period of three months.

Based on these tests, whose results are summarized in table 4, we have validated our hypothesis about entity popularity. In figure 2 we show the correlation graph between entities in the AOL logs (labeled “searched entities”) and the “named entities” inside the ANSA corpus. Often, when an entity is popular in the ANSA, it is also popular in the AOL news.

Entity Popularity Distribution. The analysis described in this section aims to find a behavioral trend in the popularity of entities. The first part of the analysis plots – for each entity type (Person, Organization, Location) – the relationship between the type and its occurrences. What is evident from Fig. 3 is that entities generally follow a common trend, with few outliers. We note that, in order for the diagram to be readable we display the logarithm of the popularity. The graph suggests the hypothesis that the ANSA person entity occurrences follow a power-law distribution. The power-law distribution implies that entities with few occurrences are very common, while entities with many occurrences are rare. For this reason, we tested if the entity occurrences were distributed

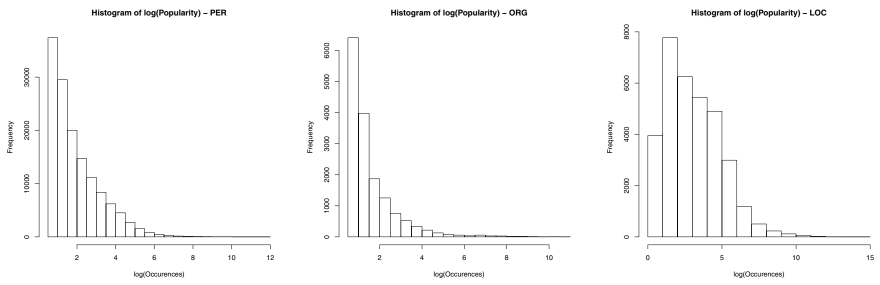

Fig. 3. Entity occurence distribution by category (logarithm): PER (left), ORG (middle), LOC (right)

according to a power-law distribution. In the past, scientists have discovered that many phenomena in the real world such as earthquakes or city growth are following a power-law pattern [19]. Detecting a true power-law distribution is however complicated [13] and requires proper tests. In the WWW, power law distributions have been used for predicting behavior on web site visits and links to pages, or for studying network links in social [18].

For calculating the power-law fitting, we follow the suggestions of Clauset et al. [6] – which are implemented in the Plfit *R* library[5] – and estimate the goodness of the fitting by using the *D* statistic of the Kolmogorov-Smirnoff test [7] (K-S test), provided by tge Plfit function. The test follows these steps:

1. First, we use the plfit library to estimate the possible parameters xmin and α where x is a vector of observations of some quantity to which we wish to fit the power-law distribution $p(x) \sim x^{-\alpha}$ for $x >= xmin$. This is done through a Maximum Likelihood Estimation process. The library returns an xmin that shows the most plausible x value, above which the power law stands for the data.
2. We use the *D* statistic of the K-S test to determine the p-value of the fit. If the p-value is over 0.05, we *cannot reject* the fact that the samples come from a power-law distribution. In this case, we will consider the power law a plausible model for tha data. Due to the fact that we do not look for the optimal model, we do not compare with other models (e.g., log-normal), but simply check the power model plausibility.

In the following paragraph we report the power-law estimation process on a sample of the ANSA dataset. The sample used for the analysis is the three month period of popularity: the same period where the experiments correlating searches to entity appearances, namely March 1st, 2006 to May 31st, 2006.

In a power-law we have $p(x) = x^{-\alpha}$. We are now able to extract the correct exponent $\alpha = 2.14$ and $xmin = 8$.

[5] R Plfit function created by Laurent Dubroca.

It appeared that, the power law distribution *is a good fit for the ANSA data.* In other words, focusing on the extracted sample data we found that a distribution $p(x) \sim x^{-2.14}, x \geq 8$ describes plausibly the empirical distribution of popularity for ANSA entities.

5 Conclusion

In this paper we have presented an approach for gaining insights into the behavior of entity popularity on the Web. We have analyzed a large corpus of news articles and their metadata, and made observations about lifespan, re-use, and popularity behavior. The latter aspect was investigated more deeply, as a first analysis had shown that popularity seems to follow a well-known distribution. To test this hypothesis, and to ensure that a generalization of the findings to the Web is viable, we have performed an additional experiment in which we correlate the entities in the news corpus with a large search engine query log. The results of the analyses are twofold. From one hand we defined entity popularity as the positive relationship between how many times an entity is mentioned and how many times is searched. On the other hand we supported this hypothesis by fitting this trend into a power-law distribution.

The results presented in this paper – in addition to their obvious contribution to "Web Science" as promoted by Tim Berners Lee et al. [2,10] – can play a role in the design of systems that deal with information about entities on the Web in various disciplines.

Acknowledgments

This work is partially supported by the by the FP7 EU Large-scale Integrating Project **OKKAM – Enabling a Web of Entities** (contract no. ICT-215032). For more details, visit `http://www.okkam.org`

References

1. Bazzanella, B., Stoermer, H., Bouquet, P.: Top Level Categories and Attributes for Entity Representation. Technical Report 1, University of Trento, Scienze della Cognizione e della Formazione (September 2008), http://eprints.biblio.unitn.it/archive/00001467/
2. Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N., Weitzner, D.J.: Creating a Science of the Web. Science 313(5788), 769–771 (2006)
3. Bouquet, P., Stoermer, H., Niederee, C., Mana, A.: Entity Name System: The Backbone of an Open and Scalable Web of Data. In: Proceedings of the IEEE International Conference on Semantic Computing, ICSC 2008, number CSS-ICSC 2008-4-28-25 in CSS-ICSC, pp. 554–561. IEEE Computer Society, Los Alamitos (August 2008)
4. Cheng, G., Ge, W., Qu, Y.: Falcons: searching and browsing entities on the semantic web. In: WWW 2008: Proceeding of the 17th International Conference on World Wide Web, pp. 1101–1102. ACM, New York (2008)

5. Cheng, T., Yan, X., Chang, K.C.-C.: Entityrank: searching entities directly and holistically. In: VLDB 2007: Proceedings of the 33rd International Conference on Very Large Data Bases. VLDB Endowment, pp. 387–398 (2007)
6. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data (2007)
7. Conover, W.J.: A Kolmogorov Goodness-of-Fit Test for Discontinuous Distributions. Journal of the American Statistical Association 67(339), 591–596 (1972)
8. International Press Telecommunications Council. Guide for implementers. Document revision 1, International Press Telecommunications Council (2009)
9. Demartini, G., Firan, C.S., Iofciu, T., Krestel, R., Nejdl, W.: A Model for Ranking Entities and Its Application to Wikipedia. In: Latin American Web Conference, LA-WEB 2008, pp. 28–38 (2008)
10. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., Weitzner, D.: Web science: an interdisciplinary approach to understanding the web. ACM Commun. 51(7), 60–69 (2008)
11. Jin, Y., Matsuo, Y., Ishizuka, M.: Ranking companies on the web using social network mining. In: Ting, H., Wu, H.-J. (eds.) Web Mining Applications in E-commerce and E-services, ch. 8, pp. 137–152. Springer, Heidelberg (2008)
12. Jin, Y., Matsuo, Y., Ishizuka, M.: Ranking entities on the web using social network mining and ranking learning. In: WWW 2008 Workshop on Social Web Search and Mining (2008)
13. Newman, M.E.J.: Power laws, pareto distributions and zipf's law. Contemporary Physics 46(5), 323–351 (2005)
14. Nie, Z., Ma, Y., Shi, S., Wen, J.-R., Ma, W.-Y.: Web object retrieval. In: WWW 2007: Proceedings of the 16th International Conference on World Wide Web, pp. 81–90. ACM, New York (2007)
15. Nie, Z., Wen, J.-R., Ma, W.-Y.: Object-level vertical search. In: CIDR, pp. 235–246 (2007), `www.crdrdb.org`
16. Pass, G., Chowdhury, A., Torgeson, C.: A picture of search. In: InfoScale 2006: Proceedings of the 1st International Conference on Scalable Information Systems. ACM Press, New York (2006)
17. Popov, B., Kitchukov, I., Angelova, K., Kiryakov, A.: Co-occurrence and Ranking of Entities. Ontotext Technology White Paper (May 2006)
18. Schnegg, M.: Reciprocity and the emergence of power laws in social networks. International Journal of Modern Physics 17(8) (August 2006)
19. Shiode, N., Batty, M.: Power law distributions in real and virtual worlds (2000)